**Exercise 4:** Similarity search with BLAST

When considering a new DNA sequence, the first question will probably be: "*What do I have in my sequence?*" which relates to the question "*Is my sequence similar to previously known sequences?*" To find out, you may compare your sequence with the public databases, using one or more computer programs that have been written for searching the databases for similarities to a given query sequence. The fastest way to search for similar sequences in the databases is **BLAST** (Basic Local Alignment Search Tool). Blast is a set of similarity search programs designed to explore all of the available [sequence databases](http://www.ncbi.nlm.nih.gov/BLAST/blast_databases.shtml) regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. Blast has several variants:

|  |  |  |
| --- | --- | --- |
| [**Sequence type**](http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/DnaTechnology?sortcol=0;table=2;up=0#sorted_table) | [**nucleotide database**](http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/DnaTechnology?sortcol=1;table=2;up=0#sorted_table) | [**amino acid database**](http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/DnaTechnology?sortcol=2;table=2;up=0#sorted_table) |
| **nucleotide query** | blastn/tblastx | blastx |
| **amino acid query** | tblastn | blastp |

* blastn is good for finding nucleotide sequences similar to yours.
* blastp is good for finding amino acid sequences similar to yours.
* blastx is good for finding amino acid sequences similar to any translation of your nucleotide sequence - e.g. if you could not recognize an ORF.
* tblastn is good for finding nucleotide sequences that can be translated into something similar to your amino acid sequence - e.g. unannotated pseudogenes.
* tblastx is good for keeping computers busy (or for very specific applications).

The BLAST 'Search' box accepts a number of different types of input and automatically determines the format. Accepted input types are: FASTA file, bare sequence or database identifiers (e.g. genbank i.d. M18533, Swissprot i.d. P11532)

* Open [BLAST](http://www.ncbi.nlm.nih.gov/BLAST) and go to *blastp (protein Blast)*
* Open the file “codes for exercises 1, 4 and 5” (on Schoology). Copy and paste the exercise 4 sequence into the sequence input field on the *blastp* page, leave the other options unchanged and press BLAST. (This sequence is translated from an mRNA sequence)
* **Which sequence in the database is most similar to the one we have submitted? Are the sequences identical?**

A Graphical overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, with the most similar hits uppermost and appearing in red. Pink, green, blue and black bars follow, representing proteins in decreasing order of similarity. Hatched areas would indicate a gap in similarity i.e., two or more distinct regions of similarity were found within the same sequence hit. Detached bars on the same line correspond to unrelated hits. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments.

* On the Blast result page click on Taxonomy reports and look at the hits for different organisms
* Go back to the Blast result page. You can access different databases by clicking on the accession code. Explore some links.
* Find the original protein, transcript variant Dp116 (NP\_004005). From which gene does this protein originate? **Find the RefSeq accession code for the corresponding mRNA sequence.**
* In the database record for NM\_004014, explore the links in the menu on the right. You can do further analysis with this sequence and link out to other databases.
* Try to find PCR primers that could be used for Dp116. There’s a link “Pick Primers” on the right side. That will take you to the Primer picking page. Let’s not worry about customizing, just scroll down to “Get Primers” and click. (This takes awhile. Select the “open in new tab” option so the wait happens in a different tab.